Center for Education Policy Research
HARVARD UNIVERSITY

STRATEGIC DATA PROJECT

SDP FELLOWSHIP
CAPSTONE REPORT 2016

Examining Next Generation Measures of Educator Effectiveness

Jack Byrd, Fort Wayne Community Schools
Laura Cain, Fort Wayne Community Schools
Kevin Eakes, Charleston County School District
Shanna Ricketts, Delaware Department of Education

Executive Summary

In recent years, the federal government has promoted and incentivized the use of student achievement measures in teacher evaluation. Responding to this challenge, Fort Wayne Community Schools, Charleston County Schools, and the Delaware Department of Education sought to learn: How reliable and valid are the components of teacher evaluation systems?

In Fort Wayne, SDP Fellows regressed student growth measures on observation in order to establish the validity of observations, revealing a positive relationship. Following the new policy in Fort Wayne that included student growth measures in teacher evaluation, the percentage of teachers rated highly effective in observations grew by 18.2%. In Charleston, an SDP Fellow analyzed change in the rater reliability of observations, comparing internal and external observer and rating the mean differences between schools. The SDP Fellow found that internal observers rated teachers higher than external observers. In Charleston, training improved reliability for teacher observations. In Delaware, an SDP Fellow evaluated the psychometric quality of educator-created assessments. The data from these assessments is intended to be used in the state's teacher evaluation system. In Delaware, their educator-created assessments proved to be reliable and valid, supporting the use of this data in teacher evaluation.

Strategic Data Project Fellowship Capstone Reports

Strategic Data Project (SDP) Fellows author capstone reports to reflect the work that they led in their education agencies during the two-year program. The reports demonstrate both the impact fellows make and the role of the SDP network in supporting their growth as data strategists. Additionally, they provide recommendations to their host agency and may serve as guides to other agencies, future fellows and researchers seeking to do similar work. *The views or opinions expressed in this report are those of the authors and do not necessarily reflect the views or position of the Center for Education Policy Research at Harvard University.*

Introduction

Over the past decade, many states and districts across the country have been rethinking their educator evaluation systems or instituting new systems. While the nuances vary, evaluation models typically include two types of metrics: classroom observations and student growth. In recent years, the federal government has promoted the revision of teacher evaluation through grant programs such as Race to the Top and the Teacher Incentive Fund that required the use of student growth as a central tenet (Lemke, Thomsen, Wayne, & Birman, 2012).

This paper includes findings from the evaluation systems in two local education agencies (LEAs)—Fort Wayne Community Schools in Indiana and Charleston County School District in South Carolina—and one state education agency (SEA), the Delaware Department of Education. Each has performed in-depth analyses investigating the reliability and validity of the different measures that make up its educator evaluation system, and is investigating relationships among the various components. Before turning to the findings of these case studies, we first summarize the relevant literature on teacher evaluation with a focus on the validity and reliability of these systems.

Literature Review

The quality of a teacher matters for student achievement (Hanushek, 2007; Koppich & Rigby, 2009; Smith & O'Day, 1990). Students of the most effective teachers have almost an entire year's worth of additional learning compared with students of the least effective teachers (Hanushek, 2007). For this reason, it is important to be able to recruit and retain high-quality teachers. But in order to do this, it is necessary to have systems in place to identify those teachers. The method for doing so—educator evaluation—continues to be both difficult (Harris, Ingle, & Rutledge, 2014) and contentious. However, strides are being made as states and districts

continue to study and refine their evaluation systems. Research has led to considerable changes in teacher evaluation resulting in rating systems in order to identify and reward quality teachers (Donaldson, 2009; Kimball & Milanowski, 2009).

Classroom Observations

Classroom observation has long been a mainstay of teacher evaluation systems. However, there has been an increase in the use of systematic tools for conducting and rating these observations—tools that define expectations and make connections to student achievement (Danielson, 2012; Gullickson, 2009). In recent years, with an increased focus on teacher effectiveness systems that include multiple measures, classroom observation instruments have proliferated, generating research about tools and data yielded from observations.

Formal observations in most systems are designed with observational components in a rubric to measure teachers' performance and instructional practices, typically resulting in a rating (Donaldson, 2009). Examining correlations of classroom observation ratings with student growth scores (e.g., value-added estimates and student learning objectives) can provide insight into the use of various components of a multiple measure teacher evaluation system. Ultimately, classroom observations are intended to be snapshots of teachers' instructional performance that, when captured multiple times across an academic year, portray typical teaching and provide opportunities for school and district leaders to support teachers to improve their instructional performance.

Challenges in Classroom Observations

The reliability of the evaluator, fostered through an executed training process, is critical to a quality evaluation. As Donaldson (2009) noted, "Evaluators need to know and be able to identify the tenets of good instruction" (p. 10). Supporting this notion, Jacob and Lefgren (2008)

acknowledged that those with an education background or training can discern quality instruction when it is observed. Still, caution must play a role: There is evidence that principals do struggle in identifying all ranges of teachers given their own biases and experiences. They asserted, "The inability of principals to distinguish between broad ranges of teacher quality suggests that one should not rely on principals for fine-grained performance determinations as might be required under certain merit pay policies" (p. 129).

The inaccuracy that surrounds teacher evaluation systems and eventual ratings continues to plague districts across the country. Now, many states and districts are required by law to develop new evaluation systems—and, in some cases, raises are tied to performance. Thus, the accuracy of evaluations is even more critical, and districts are focused on high quality professional learning for principals, as well as teachers, in the evaluation and subsequent rating of teachers.

Measures of Student Growth

Many states and/or districts have implemented models to measure a teacher's impact on student growth based on statewide standardized assessments. These models are called value-added models or student growth percentiles, but the underlying theory is the same: It is possible to statistically isolate the effect of a teacher on a student's growth from one time point to another. In addition to student growth based on statewide standardized assessments, many states and/or districts have implemented student learning objectives (SLOs) or student growth objectives (SGOs). While the specifics vary by state and/or district the premise is that, at the start of the school year, teachers set an end-of-year goal for their students. Depending on the specifics of the particular system, students may demonstrate that they have met these goals through a pre-and post-assessment, or through some other form of student achievement.

Validity and Reliability

In an educator evaluation system in which high-stakes decisions are being made as a result of the ratings an educator earns, it is important that the system be both valid and reliable. One key aspect of validity is that the inferences we intend to make based on the results or scores of assessments are subject to validation, rather than the assessments themselves. While validity and reliability are often considered with respect to assessments, one could expand the concept of validity to include not just the various components of an evaluation system, but the entire system itself. It is important to think about the inferences that are being made as a result of this system—inferences regarding the quality and effectiveness of an educator.

Messick (1995) described validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment" (p. 741). Validity has also been described as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). In other words, there is no one measure or source of evidence for validity. Rather, it is a compilation of evidence that takes many forms, such as expert judgment and relevant psychometric results. The findings presented here from three different agencies provide varying degrees of evidence for the validity of the evaluation system each has implemented.

Case Study:

Fort Wayne Community Schools

Fort Wayne Community Schools (FWCS) has been a vital part of Fort Wayne, Indiana, for more than 150 years and has a long track record of success. With 51 schools, 4,000 staff members, and 30,000 students, the district has grown along with its community. FWCS boasts an

89% graduation rate, and many alumni go on to a variety of careers including journalism, teaching, engineering, and medicine. The district takes pride in its diversity, both with respect to the student body and the offered programming. There is something at FWCS for every student, including world-class academics, championship-winning athletic teams, successful arts programs, and dozens of other clubs and activities.

FWCS is a high-poverty urban district, with 71% of students eligible for free or reduced-price lunch. The district has a very diverse ethnic population; more than half the students are non-White (Figure 1), and over 75 languages are spoken as a first language among the student population. Since 2009 there has been rapid growth in the English language learner population, which has posed challenges for the teaching staff.

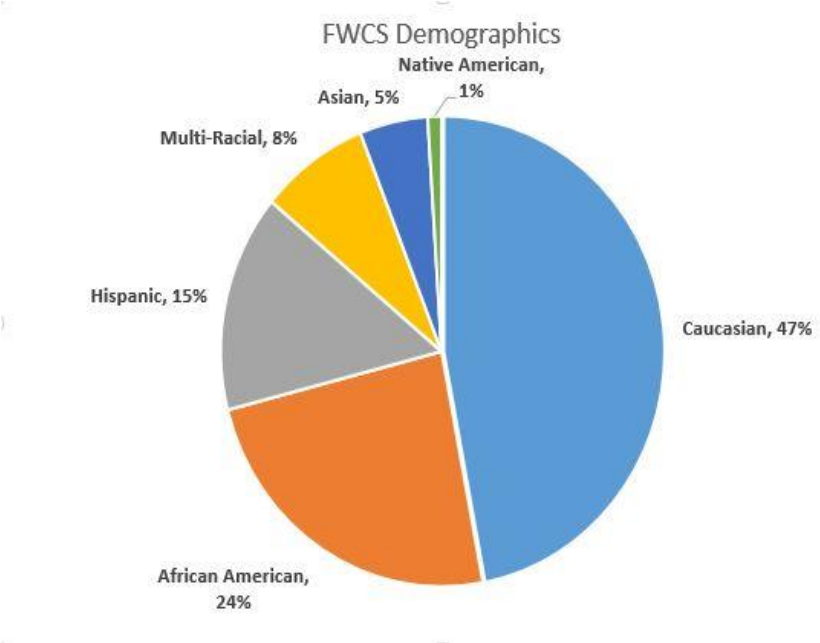


Figure 1. Fort Wayne Community Schools student racial/ethnic demographics.

Student enrollment in FWCS is strong despite increased competition and budgetary restrictions brought about by two pieces of legislation enacted by the Indiana legislature. The

first was a voucher system allowing students to attend parochial schools with free tuition. Fort Wayne has the highest percentage of students in the state taking advantage of this program. This resulted in the district's enrollment slowly declining starting in the 2012–2013 school year, averaging a loss of 300 students per year.

The second piece of legislation was enacted in 2010. Indiana constitutionally approved property tax caps, which impacted the Transportation and Capital Projects fund. By 2015, this had resulted in the need to cut the transportation budget by $2 million. FWCS was a district with school choice, but FWCS could no longer afford to provide transportation beyond neighborhood schools; this resulted in a loss of 900 students. To try to capture those students back to the district, FWCS instituted a marketing initiative to inform parents of opportunities and programs for their students. Figure 2 illustrates how these two factors have impacted FWCS enrollment.

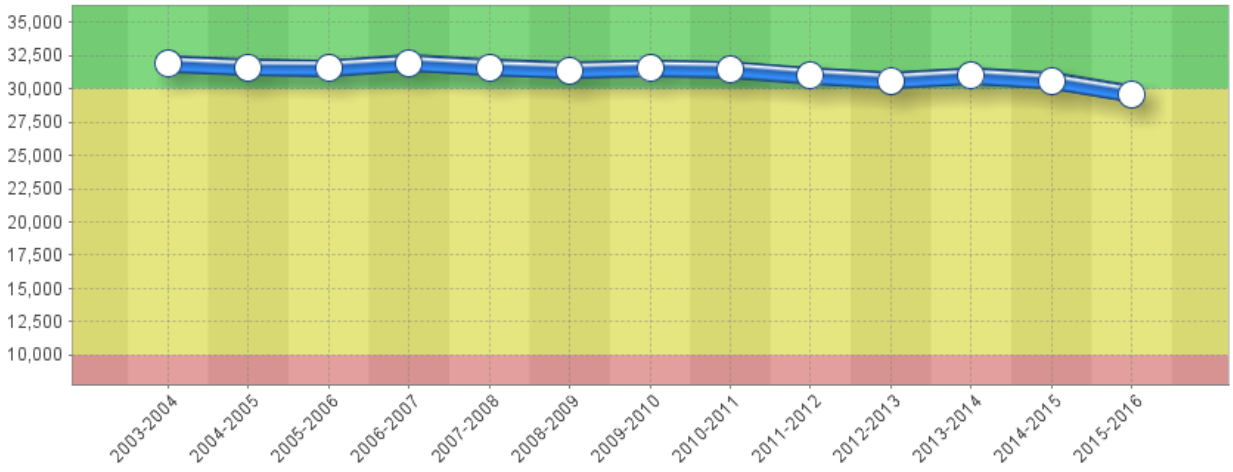


Figure 2. Impact of vouchers and transportation cuts on FWCS enrollment, 2003–2004 to 2015–2016.

Another factor that sets FWCS apart from other urban districts is the graduation rate over time (Figure 3). Despite the many challenges faced by FWCS and its students, the percentage of

students earning a high school diploma remains high. The district continues to invest in resources and programming that will support students toward on-time graduation.

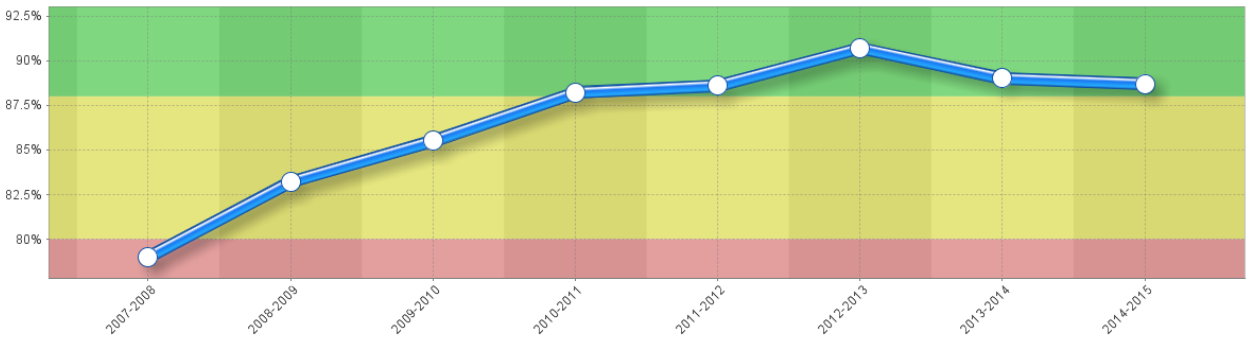


Figure 3. FWCS graduation rate over time, 2007–2008 to 2014–2015.

Policy/Research Questions

In 2010, the district instituted a turnaround model with 11 schools. This model closely followed the national model of turnaround schools with a specific element focused on teacher evaluation. The current process of evaluation was overhauled and focused instead on a system of support. Danielson (2011) described the need for ongoing, purposeful evaluation and professional learning as a priority "not because teaching is of poor quality and must be fixed but rather because teaching is so hard and it can always improve" (p. 3). This plan also included frequent and short observations. Depending on the number of classrooms within the school, these visits would take place daily or at a minimum weekly. The end result would be a final teacher rating of Highly Effective, Effective, Needs Improvement, or Ineffective. This shift in practice allowed the district to make significant changes to the evaluation process, especially after legislation was passed that required yearly teacher and principal evaluations with significant data attached to the final rating.

Professional learning with principals and assistant principals has been the priority for the last two years. FWCS focused on principals' precision in their use of the evaluation rubrics as well as the fidelity by which they were administered. Much of the work has centered on a clear and precise definition of words and phrases in the rubric and then use of those definitions to generate the observational scoring. This precise professional learning is ongoing.

To ensure that observation is reflective of student performance, school improvement performance (SIP) goals were added to the final evaluation composition. These goals made up 10% of the final score beginning with the 2012–2013 school year. Every year, educators develop academic, attendance, and disciplinary goals for their school. At the end of the school year, each goal is evaluated by indicating whether progress was made (i.e., better than the prior year), the target goal was met, or the target goal was exceeded. A final average of all goal metrics for a school is calculated as the overall SIP result. The SIP percentage of the final evaluation increased to 15% for the 2013–2014 school year.

Legislatively, all districts in the state of Indiana must attach growth data from the state assessment (ISTEP+) or other data in a significant manner. Therefore, in addition to overall SIP, beginning for the 2014–2015 school year, student growth for individual teachers was added to the final evaluation and made up 25% of the final score. Growth is determined for teachers in Grades 3–8 using the ISTEP+ assessment. Since FWCS has had a focus on literacy for all students, Dynamic Indicators of Basic Early Literacy Skills (DIBELS) or Scholastic Reading Inventory (SRI) growth is used for teachers in non-tested grades and/or subjects. For the 2014–2015 school year, SRI was added as an additional measure for teachers in Grades 3–8.

With the addition of data measures to the final evaluation, concerns were raised about the accuracy of the student growth measure and the relationship between the data measures and

observation ratings. Optimally, SIP and growth data should positively correlate with teacher observation scores. Regarding the accuracy of student growth, concerns were expressed about the targets being used to measure growth and the scale being used to score the growth. As a result, adjustments were made over the 2-year period. Therefore, a study was commissioned to answer the following research questions:

1. What is the trend of components used for the final ratings (i.e., observation, SIP, and student growth)?
2. What has been the impact of the student growth calculation changes?
3. How has the addition of data measures affected the final rating? Is one data measure impacting the final rating more than another?
4. Is there a correlation or an increasing correlation between the data measures and observation ratings?

Findings

Over the past four years, final observational scores have grown significantly. The percentage of teachers observed as Highly Effective grew 9.3 percentage points from 2011–2012 to 2012–2013. After remaining fairly constant for 2012–2013 and 2013–2014, this percentage grew another 8.4 points for the 2014–2015 school year. This represents an 18.2% increase in teachers rated Highly Effective. Put another way, approximately 364 more teachers were observed as Highly Effective in the 2014–2015 school year than in the initial year of 2011–2012 (Figure 4).

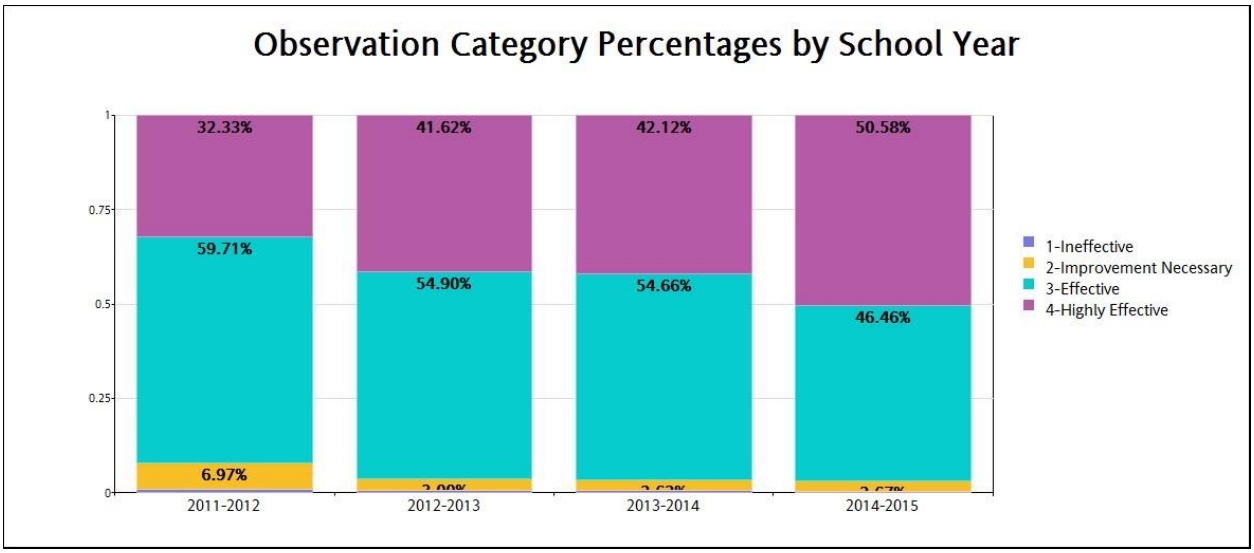


Figure 4. Observation category percentages by school year, 2011–2012 to 2014–2015.

Once the data components were added to the observation score, however, a reverse effect for Highly Effective teachers was observed. Due to the different final rating composition, only the 2013–2014 and 2014–2015 school years were used for this comparison. For these school years, observation was 60%, SIP was 15%, and student growth was 25% of the final rating. While the percentage of teachers receiving a final category rating of Effective or Highly Effective increased by 1.86 percentage points, the percentage of teachers in the Highly Effective category decreased by almost 6% (Figure 5).

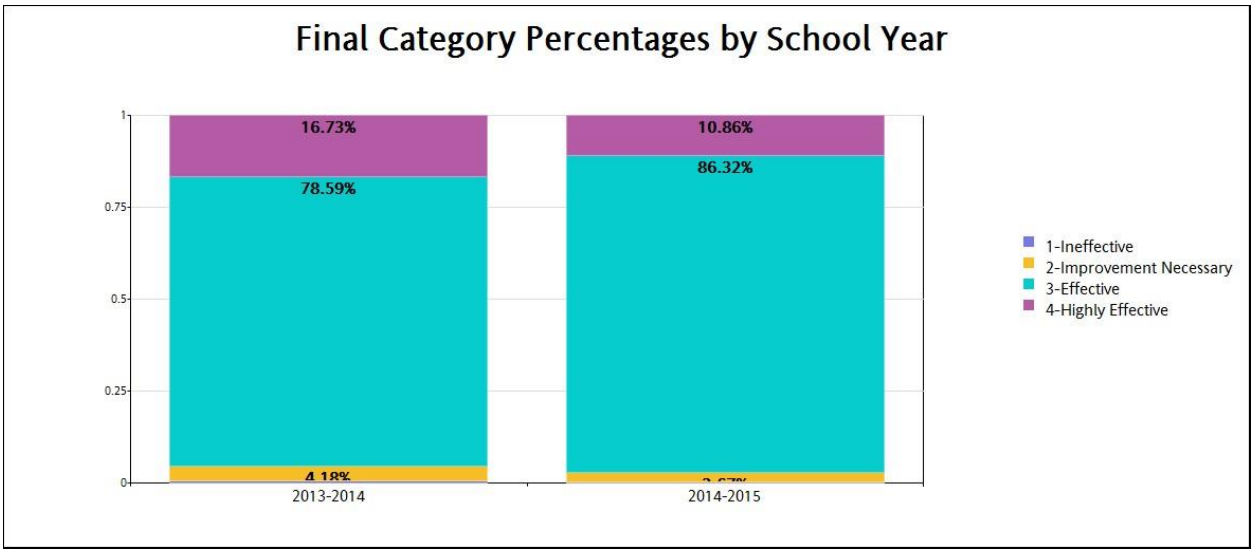


Figure 5. Final observation category percentages by school year, 2013–2014 to 2014–2015.

This decrease in the percentage of highly effective teachers underlies the concerns expressed about the application of data. Therefore, each component of the final rating was examined for both years. It must be noted, however, that the student growth metric underwent modifications between 2013–2014 and 2014–2015. Specifically, there were three major changes to how student growth was calculated. First, for 2014–2015, growth was calculated using a rubric scoring system as opposed to the scale system used in 2013–2014. Second, for the 2013–2014 school year, teachers who were reported to the State of Indiana Department of Education (DOE) as responsible for an ISTEP+ content area (i.e., math or language) used the DOE calculated teacher growth metric; those teachers not reported to the DOE used a growth metric calculated as the mean growth of their rostered students enrolled in their classes 90% of the class time. For 2014–2015, all teachers used a growth metric calculated as the mean growth of their students enrolled 90% of the class time. And third, for 2013–2014, ISTEP+ growth was calculated as median growth, while for 2014–2015 it was calculated as mean growth. The primary reason for this change was that the DOE growth metric used for teachers in 2013–2014 used median student growth, and comparable calculations were performed for those teachers

with ISTEP+ growth but not reported to the DOE. For 2014–2015, the mean was used since the DOE growth metric was no longer utilized.

In order to create equitable answer sets, only teachers whose ratings utilized class-based growth were used in the comparison and the 2014–2015 student growth was scored using the 2013–2014 scale. Both years used mean student growth. Figure 6 clearly shows that the changes made to the growth calculation for 2014–2015 were to the advantage of the teacher (student growth)—when student growth was standardized (standard growth), there was insignificant change between school years. As expected, observation scores increased; however, average SIP scores decreased by .57 points. To ensure this was not a condition limited to this subset of teachers, the average SIP score for all teachers was calculated, resulting in a .53-point decrease between school years.

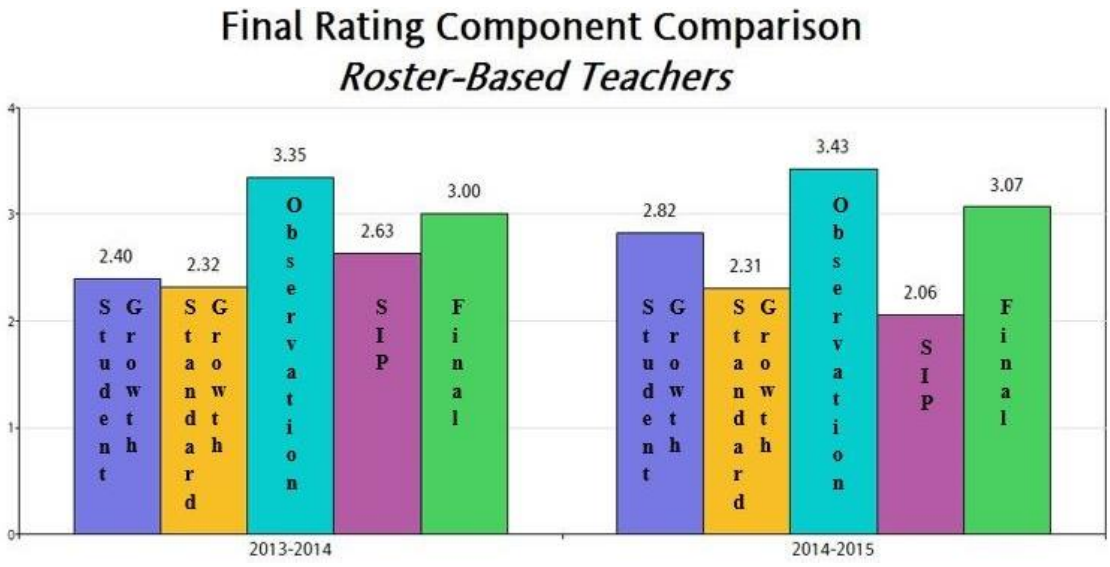


Figure 6. Final rating component comparison of roster-based teachers. Rating category point ranges are as follows: Highly Effective = 4.0–3.5, Effective = 3.49–2.5, Improvement Necessary = 2.49–1.75, and Ineffective < 1.75.

It is interesting to note that while the percentage of highly effective teachers decreased for the 2014–2015 school year, the average final rating increased for the subset population. For all teachers, the average final rating also increased (from 3.06 to 3.11). A boxplot and histogram (Figure 7) shows the distribution of final scores. The teacher scores were more closely grouped (.04117 SD in 2013–2014 versus .3279 SD in 2014–2015). It is hoped that this closer grouping is due to evaluators observing teachers more consistently.

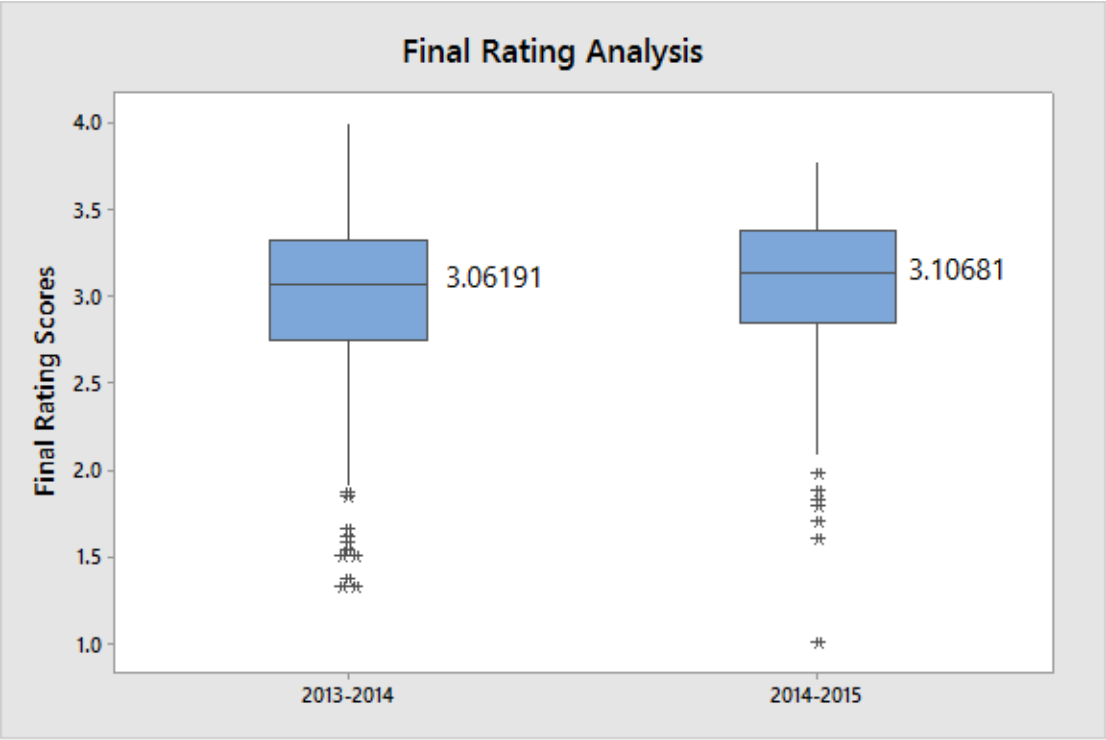


Figure 7. Comparison of final ratings for 2013–2014 and 2014–2015. Rating category point ranges are as follows: Highly Effective = 4.0–3.5, Effective = 3.49–2.5, Improvement Necessary = 2.49–1.75, and Ineffective < 1.75.

Finally, the relationship between student growth and observation scores was examined for 2013–2014 and 2014–2015 (see Figures 8 and 9, respectively). For both school years, growth was not a major predictor of observation scores nor was there a strong correlation—less so in 2014–2015.

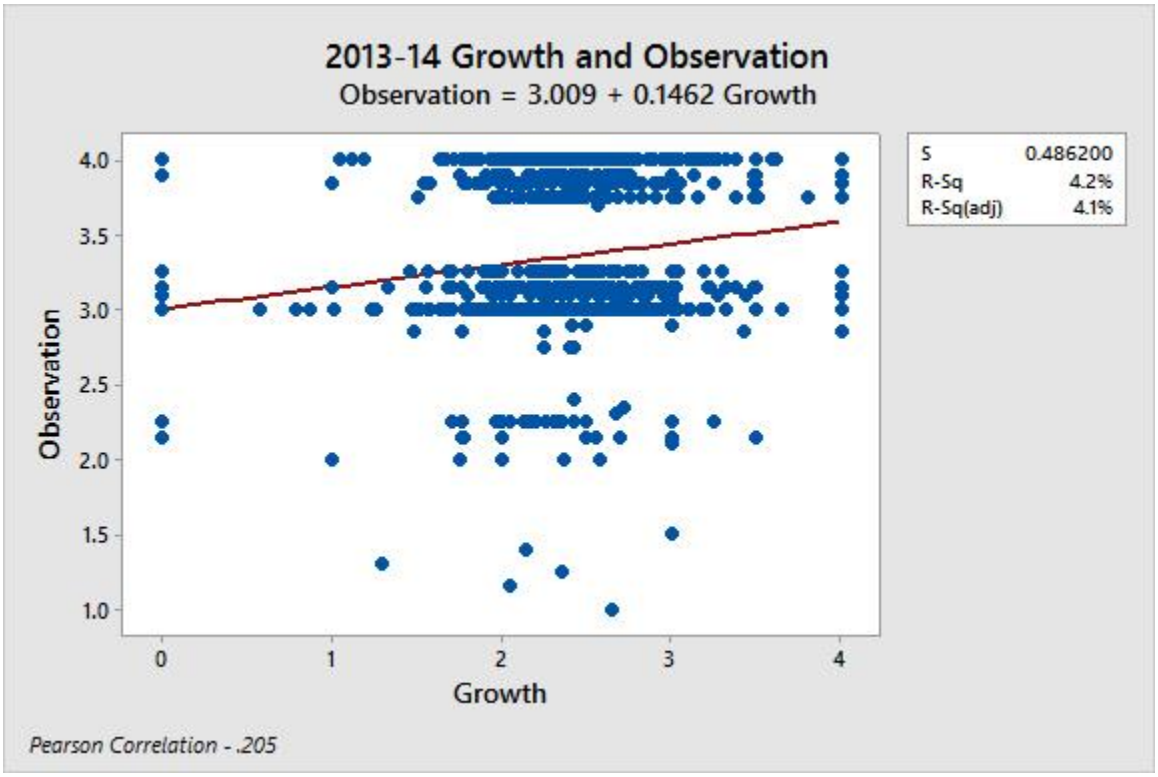


Figure 8. 2013–2014 growth and observation. Rating category point ranges are as follows: Highly Effective = 4.0–3.5, Effective = 3.49–2.5, Improvement Necessary = 2.49–1.75, and Ineffective < 1.75.

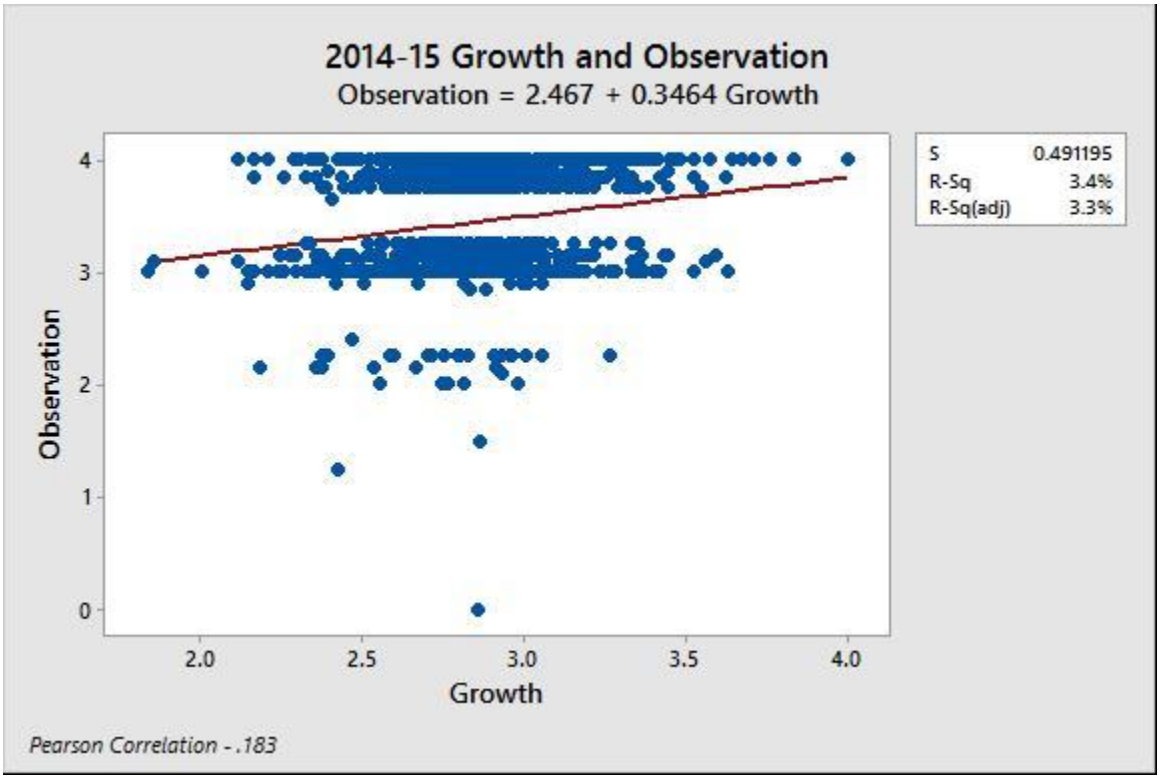


Figure 9. 2014–2015 growth and observation. Rating category point ranges are as follows: Highly Effective = 4.0–3.5, Effective = 3.49–2.5, Improvement Necessary = 2.49–1.75, and Ineffective < 1.75.

Conclusion

Trend of components used for final ratings. Observation ratings increased significantly (9.3%) between the 2013–2014 and 2014–2015 school years and were grouped closer together. There are several reasons this could have occurred. First, evaluators could have been compensating for the addition of the data measures. Principals were well aware of the impact of data on final teacher ratings, and we know that the principal is a distinct variable in the reliability of the observational score. This particular data point requires that the district spend time and effort on professional learning for principals and those who supervise principals in order to monitor reliability of the process.

Impact of student growth calculation changes. The adjustments made to the student growth calculation were in response to a concern of the teachers association. They felt the original calculation was too stringent (i.e., targets were too narrow) and needed to be placed on the same scale as the observation rating. When changes were made for 2014–2015 school year, student growth scores increased. However, to determine whether student performance also increased, student growth would need to be calculated for the two school years using the same model. When student growth is standardized for both years, it remains flat. This means the increase in student growth was only mathematical, not due an increase in student performance.

Effect of data measures on final ratings. The addition of the data measures resulted in more teachers from the Highly Effective observation category ending in the Effective final category. More specifically, for the 2013–2014 school year, 60% of the Highly Effective observed teachers moved to the Effective final category, while for the 2014–2015 school year, 79% moved. Only a very small percentage of teachers (3%) moved from an Effective observed category to an Improvement Necessary final category. This impact caused much discussion regarding the added data, especially for teachers who, after their observational conference, believed they would have a Highly Effective rating. This perception caused some anxiety when their final rating was calculated and it prompted the above concern regarding the growth calculation. It is also important to note that a stipend was given to teachers rated Highly Effective and Effective. The difference in dollars between the two for the 2014–2015 school year was $500 per teacher.

Relative impact of data measures on final rating. With the above finding, even with the increase in student growth ratings, more teachers ended in a lower final category then they may have anticipated in looking at their observational scores. In looking at the different

components, it was found that the SIP goals with target attainment impacted the overall ratings more than other components of the rating calculation.

Correlation between data measures and observation ratings. There was little correlation between observation and student growth scores, and it decreased for the 2014–2015 school year. The relationship between the two factors was also very slight. This particular element is troubling to the district, as student growth is always one of their priorities, especially in the area of literacy. The fact that the observational score had little correlation to the achievement data supports a large need to better understand and recognize rigorous instruction and work with principals and teachers to improve the overall performance of teachers and students.

Next Steps

Consistently measure student growth. In order for more meaningful conclusions to be obtained from the data, the student growth calculation as it applies to the final rating needs to stabilize and be consistent for several years. This will allow us to verify that the correlation with the observation data is accurate.

Revise SIP process. SIP results showed the largest decrease of all components and had the greatest negative effect on teacher final ratings. The district needs to review the process of SIP goal development and continue to provide support for those schools that are consistently not meeting targets.

Continue to research ways to effectively deliver literacy instruction in all content areas. Since literacy data are applied across all teaching levels and content areas, further research needs to occur regarding teacher disciplines and success of student growth. The district needs to know more about teachers' ability to provide adequate literacy instruction across all

disciplines, what best practices already exist in the district, and how to replicate those in more classrooms.

Continue efforts to train evaluators on proper observation practices. While the consistency of observations has increased, the relationship to student growth has decreased. Evaluators need to be able to properly identify ineffective instructional practices, recommend steps for improvement, and objectively create summative evaluations that promote both teacher effectiveness and student achievement.

Case Study:

Charleston County School District

Charleston County School District (CCSD) is the second largest school system in South Carolina, representing a unique blend of urban, suburban, and rural schools that span 1,000 square miles of coastal lands. CCSD serves nearly 50,000 students in 86 schools and specialized programs. The district offers an expanding portfolio of options—including neighborhood, charter, magnet, international baccalaureate, and Montessori schools—and is divided into elementary, middle, and secondary learning communities. More than 3,500 teachers serve a diverse student population that is 46.6% White, 40.1% Black, 8.6% Hispanic, 3.2% multiracial/other, and 1.5% Asian. Approximately 7% of CCSD's students are English language learners, nearly 9% percent of students receive special education, and 50% of students qualify for free or reduced-price lunch.

Policy/Research Questions

CCSD received a federal Teacher Incentive Fund (TIF) grant in October 2012. A primary focus of the grant was to create and implement a multiple measure educator evaluation system with student growth as a significant component of the model. The evaluation system was to

differentiate educator effectiveness across several levels, which was a departure from CCSD's previous dichotomous evaluation system that yielded outcomes of either Met or Not Met. Teacher and principal evaluation models were created through a process of soliciting input from educator work groups and external consultant partners and then receiving approval from school district leadership and the federal grant office. The focus of the current project was the multiple measure teacher evaluation model.

CCSD's teacher evaluation model consisted of three components: classroom observations, student growth (i.e., value-added estimates and SLOs), and state evaluation. Scores from individual components were combined to generate an overall teacher effectiveness score ranging from Highly Effective to Ineffective. South Carolina required that CCSD continue to report state evaluation results and that the outcome for each teacher was the final metric of the model. However, it is excluded from this project.

The Strategic Data Project (SDP) Fellow enrolled from CCSD investigated the following aspects of the district's new teacher evaluation model:

1. How can classroom observation reliability be improved?
2. How do classroom observation ratings compare between different types of observers?
3. How do classroom observation ratings compare based on schools' poverty levels?
4. What relationships exist between value-added estimates and SLO scores?
5. What relationships exist between classroom observations and student growth scores?
6. How do teachers' overall effectiveness scores vary across years?

Project data were gathered during the 2013–2014 and 2014–2015 school years. As shown in Figure 10, CCSD's teacher evaluation model consisted of three components, weighted as indicated: classroom observations (35%), student growth including value-added estimates and

SLE scores (35%), and ADEPT, the state evaluation (30%). Overall teacher effectiveness averages were also examined. In the first year, teachers at 14 pilot schools participated in the new evaluation model; in the second year, the classroom observation component was expanded to all schools and teachers in the district.

Figure 10. CCSD's teacher evaluation model.

The district-designed classroom observation tool (COT) was the instrument used to measure classroom performance. The COT is based on South Carolina's teaching standards, which are rooted in the Charlotte Danielson framework. The COT was developed during a year-long process with focus group input from teachers, principals, and curriculum specialists. Evaluators rated teacher classroom performance during unannounced observations of full-length lessons, using 18 criteria with a scale from 1 to 4 corresponding to Ineffective, Minimally

Effective, Effective, and Highly Effective, respectively. Ratings from each of the 18 criteria were averaged to produce an overall score per observation. To avoid a "checklist approach," focus groups developed a rubric that detailed teaching performance for all 18 criteria across all rating categories with footnoted grade-level and subject-specific examples for evaluators to consider. The COT required numerical ratings and narrative feedback for each teacher from school-based administrators (i.e., internal evaluators) as well as central office staff (i.e., external evaluators). Teachers also had the option to submit lesson reflections to the observers to inform ratings and recommendations.

To address rater reliability, evaluators were required to engage in a specific training and observation certification process prior to becoming eligible to complete COTs. Potential evaluators participated in school-based COT sessions during which teams of evaluators visited classrooms, rated teachers, and discussed ratings. Additionally, professional development sessions were provided on the COT during administrator meetings. Evaluators were required to pass a district-created exam that included rating a recorded lesson and responding to items about COT protocol with 80% accuracy or better. Value-added estimates provided one measure of student growth for this project. Teachers of students in Grades 4–8 in math, science, social studies, and language arts, as well as high school teachers of students in English I, Algebra I, and Biology, earned value-added estimates. Teachers completed roster verification to ensure accurate linkage and attributable instructional responsibility between teachers and students. CCSD's value-added model includes a recent test score and covariates to control for factors such as student attendance, special education status, English language learner status, student mobility during the school year, and students who are over age relative to their grade level.

Data from SLOs provided student growth measures for teachers of students in non-tested subjects. Over 200 SLOs were created by CCSD teachers and curriculum specialists and were prescriptive in design to promote consistency across the district. SLOs for each grade/subject/course used the same assessment measures, which included available tests already in place (e.g., Measures of Academic Progress) and district-created assessments. Additionally, most SLOs included a performance task that was scored using a rubric. The district set specific guidelines for determining growth targets for each assessment. Overall teacher effectiveness ratings were computed from the combination of all component scores. Corresponding ratings for overall scores are provided in Table 1.

Table 1

Ratings Assigned to Overall Teacher Effectiveness Scores

Rating	Category
4.00–3.60	Highly Effective
3.59–2.80	Effective
2.79–2.00	Minimally Effective
1.99–1.00	Ineffective

Findings

Classroom observation rater reliability. When the COT was introduced, rater reliability was low ($\alpha = .38$). It was evident that even with a well-defined rubric, evaluators' perceptions of what constituted a rating of Effective varied widely. Through training and increased familiarity with the instrument, rater reliability steadily improved to a high of .94 for the most recent school year, when COTs were required in all CCSD schools for all teachers. Evaluators attributed improvements in COT scoring consistency to ongoing training and the certification process.

Classroom observation score comparison by evaluator type. For this project, 9,273 completed COTs were examined. Internal observers rated teachers more highly (M = 3.13, SD = 0.40) than external observers (M = 3.07, SD = .37); the difference was significant, t (9271) = 6.419, $p < .001$, yet effect size was relatively small (d = .16). Higher observation scores from site-based administrators than from outside observers has been noted in CCSD across multiple years and has been empirically documented. The use of external, objective observers mitigates potentially inflated scores from site-based administrators who may have a subjective perspective. Indeed, observers commented that knowing other professionals would be providing ratings for the same teacher led them to attend closely to the rubric. Some teachers positively noted that having multiple observers with different lenses provided validity to their ratings and stated that a single observation and single observer would not afford the same opportunity. The use of internal and external observers was valuable and will continue.

Classroom observation score comparison by school poverty status. Schools were classified as high poverty or non-high poverty based on federal guidelines for Title I eligibility. Evaluators rated teachers at non-high poverty schools higher (M = 3.20, SD = .35) than teachers at high poverty schools (M = 3.03, SD = .41); the difference was significant, t (9271) = -21.195, $p < .001$, with moderate effect size (d = .45). This finding is consistent with other studies, suggesting that some school and student characteristics may influence teacher observation scores (Chaplin, Gill, Thompkins, & Miller, 2014; Whitehurst, Chingos, & Lindquist, 2015). Specifically, the largest scoring gaps were found in the four areas highlighted in Figure 11.

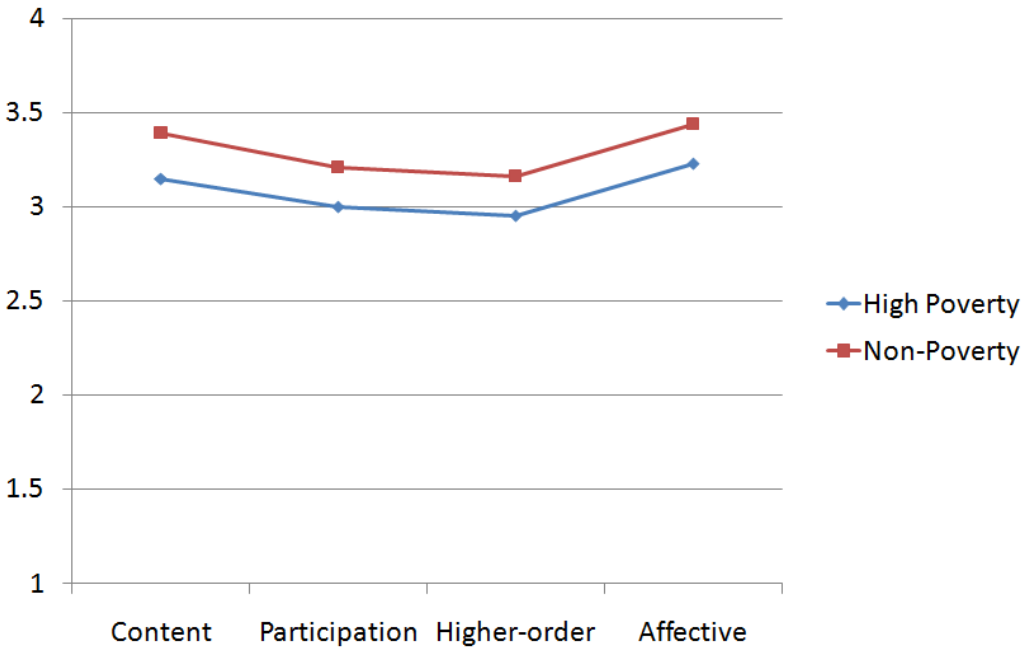


Figure 11. Largest scoring gaps in classroom observations by poverty status of school.

Relationship of scores from value-added estimates and student learning objectives. Value-added and SLO scores from teachers with both student growth measures were compared (n = 105). The sample included teachers from Grades 4–8 in state-assessed subjects and high school teachers of students with state-required end-of-course tests. Scores from SLOs were higher (M = 2.89, SD = .84) than value-added scores (M = 2.78, SD = .58). In the aggregate, there was a small positive correlation of 0.16 between value-added and SLO scores that was not significant. Similarly, when disaggregated, there was not a significant correlation between high school value-added estimates and SLO scores. A moderate, positive correlation of .38 was found between value-added estimates and SLO scores for teachers of state-tested subjects in Grades 4–8 (n = 52) that was significant (p < .01). The measures used for SLOs in Grades 4–8 relied on available testing formats (e.g., Measures of Academic Progress, DIBELS), while secondary SLOs relied on district-created assessments that had yet to be vetted for reliability.

Relationship of scores from classroom observations and student growth measures. Relationships between classroom observation scores and both types of student growth measures were examined. For the 125 teachers with classroom observation scores and a value-added estimate, a moderate, positive correlation of .29 was found that was significant ($p < .01$). Similarly, for the 105 teachers with classroom observation scores and an SLO score, a slightly smaller positive correlation of .23 was found that was significant ($p < .05$). Classroom observation scores were higher than both value-added estimates and SLO scores.

Overall teacher effectiveness scores show variability across years. For the two years in which overall teacher effectiveness scores were available, an upward trend in these ratings was observed. In 2013–2014, the overall effectiveness average was 3.08; in 2014–2015, the overall average was 3.23. Potential factors relevant to overall score improvements were familiarity with the evaluation system and the reality that during the second year four times the number of teachers were evaluated compared to the first year.

Still, more teachers improved than regressed with regard to scores from the first to second year. Of teachers scoring Effective or higher during 2013–2014, 51% earned increased overall effectiveness scores during 2014–2015, while 14% of teachers scoring Effective in the first year earned lower scores during the second year, with two teachers moving below Effective. Of teachers scoring below Effective during 2013–2014, 39% earned increased overall effectiveness scores during 2014–2015 and moved into the Effective category.

Next Steps

CCSD's new teacher evaluation model has provided more data about every teacher in the district—and about student growth—than the district has previously enjoyed. Analysis from this project informed the following recommendation to guide future work:

- Continually monitor rater reliability and design a system to identify evaluators with scoring inconsistencies so that assistance can be provided.
- Qualitatively analyze evaluators' feedback on the classroom observation tool, and create professional learning opportunities to enhance feedback for teachers.
- Maintain multiple perspectives (e.g., internal and external evaluators) for classroom observations.
- Explore the use of technology for observations, since the TIF grant is sun-setting and financial resources for external evaluators will diminish.
- Expand teacher observations to include self-assessment through the use of technology and online professional learning platforms.
- Complete additional analyses of differences in classroom observation scores at low and high poverty schools, compare these scores with student growth and proficiency data, and determine the need for additional training for observing in varied classroom settings.
- Conduct item analysis and reliability studies for district-created SLO assessments; refine these with each administration.
- Design a system to monitor ratings provided on SLO performance tasks to promote consistency across the district.
- Provide professional learning for teachers and principals on scoring SLO performance tasks, which will also improve classroom formative assessment.
- Consider expanding the teacher evaluation model to include student voice surveys.
- Continue to examine overall teacher effectiveness scores longitudinally to determine the most salient factors to include in future educator evaluation models.

Case Study:

Delaware Department of Education

The Delaware Department of Education (DDOE) has 19 traditional school districts and 26 charter school districts. About 10,000 teachers and specialists serve approximately 140,000 students. Delaware's student population is 48% White, 31% Black, and 15% Hispanic. Delaware's mission is that every student will graduate from high school ready for both college and career.

Policy/Research Questions

Delaware's revised evaluation system, the Delaware Performance Appraisal System II (DPAS-II), has been in place since the 2011–2012 school year. In addition to the four observational components, a student improvement component (Component V) was also included. See Figure 12.

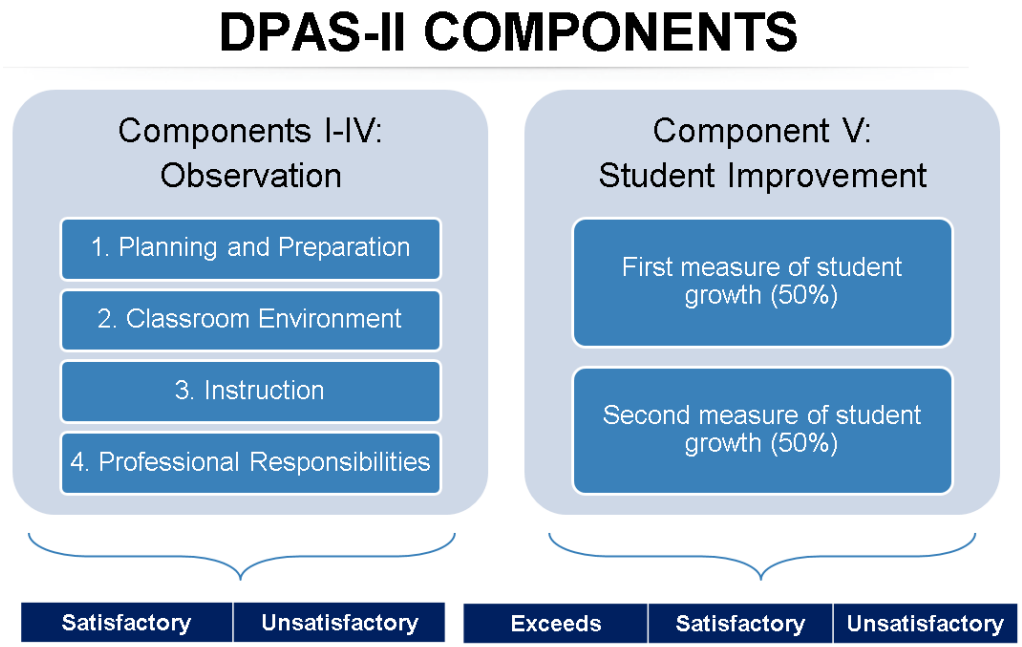


Figure 12. Components I–V of the Delaware Performance Appraisal System II.

Component V consists of three measures, as follows:

- Measure A—State assessment. This was the Delaware CAS as of the 2014–2015 school year; it has been replaced by Smarter.
- Measure B—Bank of pre- and post-assessments.
- Measure C—Bank of student learning objectives (SLOs) and student growth objectives (SGOs).

Measure B comprises external, internal, and alternate assessments (Figure 13). Over 200 pre- and post-assessments make up the internal Measure B assessments. These student growth measures were created by Delaware educators in collaboration with content experts at the DOE and external consultants who ensured quality control and provided technical oversight. During the first three years of implementation there was discussion throughout the state—as evidenced by annual statewide survey results and meeting minutes from state advisory committees—about the validity and reliability of internal Measure Bs. Thus, they are the focus of this study.

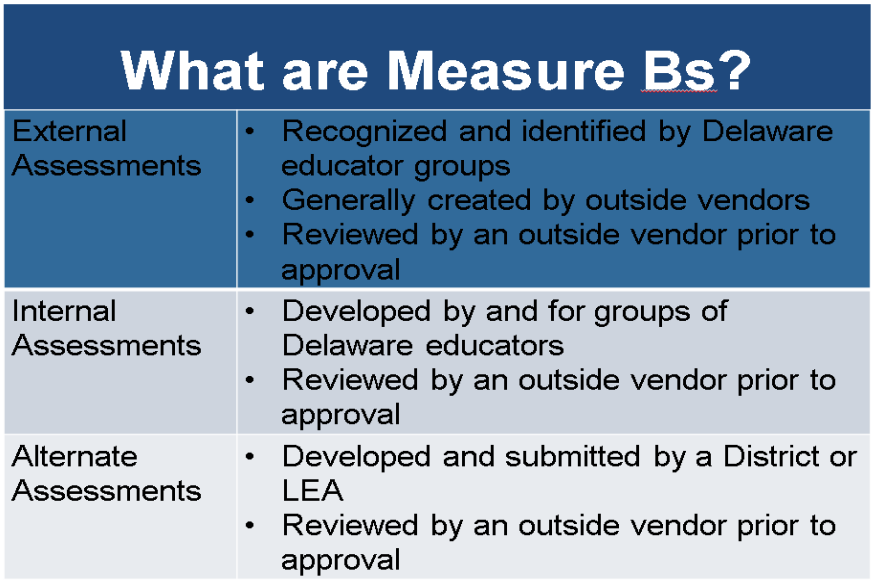


Figure 13. Aspects of DPAS-II, Component V, Measure B.

The internal Measure B assessments were designed to be aligned to content being taught within classrooms and to allow educators in Delaware to measure student academic growth in almost every subject area over the course of the school year. Some of the most used assessments are Social Studies Grade 8, Science Grade 8, and ELA Grade 10; some of the least-used assessments are German Level III, Sheet Metal I, and Textiles and Clothing III. The two research questions that this sought to answer were:

1. What is the psychometric quality of the educator-created assessments used as part of Delaware's educator evaluation system?
2. How can the results of past student performance on these assessments be used to inform the goal-setting process for both teachers and administrators?

Findings

Now that the early research has been conducted, this case study provides evidence for the validity (and thus reliability) of these assessments. The vast majority of Delaware's assessments exhibited adequate to good reliability. Further, there was a great amount of evidence for the validity of the assessments for measuring both student growth and educator effectiveness.

Evidence for validity in the construction of the assessments. As noted by Herman, Heritage, and Goldschmidt (2011), and building on Kane's (2013) claim framework, a key component of establishing validity is identifying the purpose of the assessment. As part of DPAS II, Measure B assessments are intended to demonstrate, in aggregate, teachers' contributions to their students' learning. Specifically, the assessments should provide evidence of content mastery for relevant standards and, importantly, be amenable to demonstrating growth—which also implies that they are sensitive to the quality, consistency, and rigor of instruction.

As described above, these assessments (part of Delaware's library of student growth measures) were designed by Delaware educators for almost every grade and subject area statewide. Assessments are written to a set of standards that align with the content that students are expected to learn in an academic year. Educators and content experts from the DDOE and across the country partner to write items that will be used on the assessment. An external vendor provides assessment-writing advice, training, and support to educators throughout this process. The vendor then assures the quality of the assessments by ensuring that they exhibit appropriate properties, including alignment of items to the standards, unambiguous items, an appropriate scoring format, adequate representation of the intended domain, and an appropriate mix of item difficulty. The Delaware Measure B process is substantively grounded in the expertise of educators and moderated by additional independent expertise, which is a solid basis for establishing validity.

Evidence of fairness. Validity also relates to evidence of fairness. In Delaware's educator evaluation system, fairness relates to two aspects: (a) the appropriateness of the assessment for students and their opportunities to learn the material that have been provided them, and (b) whether performance reflects what the standards denote (and thus what teachers teach). Analyses performed by DDOE show that students consistently perform better on Measure B posttests (after instruction) than on Measure B pretests (upon entering a class).

Psychometric quality of the assessments. The posttests from the 2013–2014 school year were examined in great depth to give DDOE a fuller understanding of how well the state-approved assessments were performing psychometrically. Measure Bs were examined for reliability as measured by Cronbach's alpha, a measure of internal consistency. The reliability of

the majority of the assessments was acceptable (generally considered to be a Cronbach's alpha > 0.7). Table 2 gives a snapshot of the reliability estimates of some of the most-used assessments.

Table 2

Reliability of Key Measure B Assessments

Assessment	Reliability
Social Studies (Grade 8)	0.76
Science (Biology, Grade 10)	0.82
Social Studies (Grade 7)	0.79
Science (Grade 8)	0.83
Science (Grade 7)	0.82
WLD (Spanish, Level I)	0.39
Science (Grade 6)	0.86
Science (Earth and Physical Sciences, Grade 9)	0.83
Social Studies (Grade 6)	0.81
Mathematics (Grade 6)	0.72

Some assessments did not exhibit high internal consistency. However, further examination revealed that this may have had more to do with the structure of those assessments than with the items in general. For example, assessments with a heavily-weighted constructed response section tended to exhibit lower reliability than assessments that used entirely multiple choice responses. The DDOE also examined other relevant psychometric indicators, such as differential item functioning (whether items were equally difficult for subsets of students), range of difficulty (whether each Measure B had an adequate distribution of easy, moderate, and hard items), discrimination and fit (whether items were logically related to the assessments and the skill levels of students taking it, and whether there were any outliers).

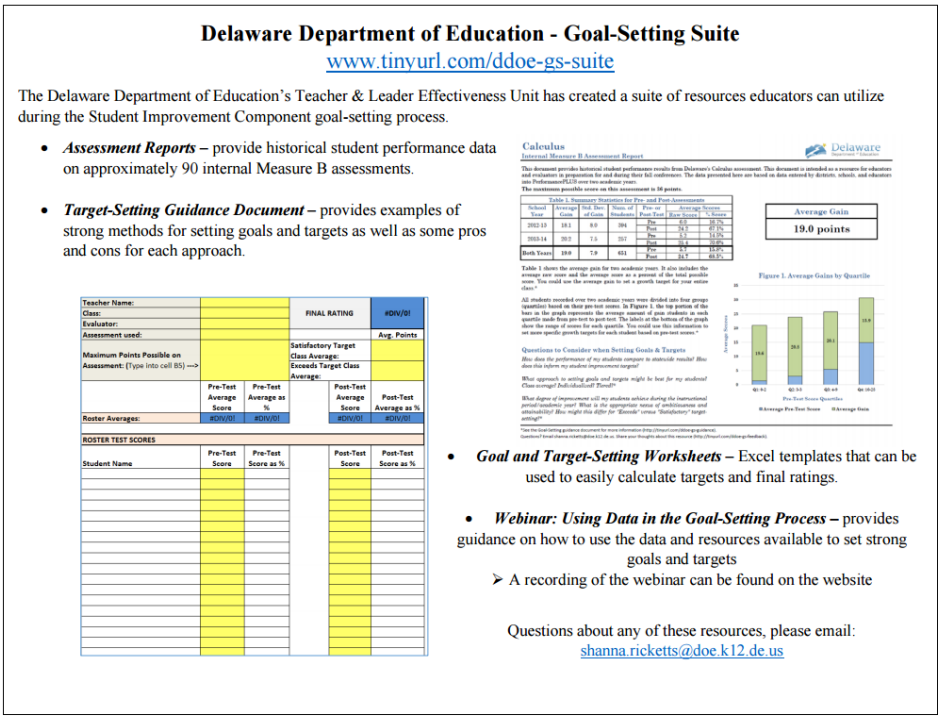
Convergent evidence of validity. Another type of evidence for validity of an assessment is convergent evidence. This refers to whether there is a relationship between a Measure B assessment and another assessment intended to measure the same construct (AERA, APA, & NCME, 2014). The Measure B assessments were highly correlated with other assessments that

measured similar constructs. For example, student scores on the mathematics Measure B internal assessments were highly correlated with student scores on the mathematics DCAS assessment for that same year. If both assessments were measuring the same construct (in this case, mathematics student growth) then there should be a high correlation between scores on each of these assessments. This was found to be the case with correlations around 0.7 or higher, thus providing convergent evidence of validity.

Validity in the use of the assessments (consequential evidence). While the vast majority of the assessments themselves exhibited average or good reliability, in and of itself this is not sufficient to guarantee the validity of the assessment. In order to explore the validity of the assessments, one also has to identify how they will be used: "Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use" (AERA, APA, & NCME, 2014, p. 11). The claim based on Measure B results is that the change in student scores between the pretest and the posttest reflects a change in student skills and knowledge related to a particular subject, and that this change in skills and knowledge was facilitated by the student's teacher in that class.

This is relevant because DPAS II declares that effective teacher practices are evidenced by multiple indicators; demonstrated teacher effectiveness at providing opportunities for students to learn the subject matter is one important aspect. Technically, the reliability of the assessments is sufficient to allow for inferences about mean changes in pretest and posttest performance. Equally important are the consequences related to the inferences. Claims about teachers based on Measure B results are vetted by principals and provide only one of several indicators of a teacher's effectiveness. This line of consequences is thus appropriately matched to the nature of the evidence.

Using assessment results.



Delaware Department of Education - Goal-Setting Suite

www.tinyurl.com/ddoe-gs-suite

The Delaware Department of Education's Teacher & Leader Effectiveness Unit has created a suite of resources educators can utilize during the Student Improvement Component goal-setting process.

- ***Assessment Reports*** **–** provide historical student performance data on approximately 90 internal Measure B assessments.
- ***Target-Setting Guidance Document*** **–** provides examples of strong methods for setting goals and targets as well as some pros and cons for each approach.

Teacher Name:		FINAL RATING	#DIV/0!
Class:			
Evaluator:			Avg. Points
Assessment used:			
Maximum Points Possible on Assessment: (Type into cell B5) --->		Satisfactory Target Class Average:	
		Exceeds Target Class Average:	
	Pre-Test Average Score	Pre-Test Average as %	Post-Test Average Score	Post-Test Average as %
Roster Averages:	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!

ROSTER TEST SCORES				
Student Name	Pre-Test Score	Pre-Test Score as %	Post-Test Score	Post-Test Score as %

- ***Goal and Target-Setting Worksheets*** **–** Excel templates that can be used to easily calculate targets and final ratings.
- ***Webinar: Using Data in the Goal-Setting Process*** **–** provides guidance on how to use the data and resources available to set strong goals and targets
  - ➢ A recording of the webinar can be found on the website

Questions about any of these resources, please email:
shanna.ricketts@doe.k12.de.us

Next Steps

Maintain and improve Measure B assessments. In addition to providing for quality control at the front end of the assessment creation process, Delaware has implemented a process referred to as the "refinement cycle" (or "wash cycle"). This process dictates that 25% of the assessments come up for review by DDOE, alongside educators, each year. When an assessment comes up for review, Delaware educators take an in-depth look at it and revise or remove items that have not performed well historically. In the spirit of continuous improvement, DDOE continues to extensively monitor these assessments and to identify ways in which they can be improved, taking into account both the technical properties that are needed for a quality assessment, as well as input from Delaware educators regarding the alignment of standards and content covered on the assessment to the standards and content taught in Delaware classrooms.

Lessons Learned

Fort Wayne Community Schools (FWCS), the Charleston County School District (CCSD), and the Delaware Department of Education (DDOE) continue to focus on improving educator evaluation. Requirements from state and federal entities mandating the use of student growth in teacher evaluation provided part of the impetus to engage in this work. Equally motivating was the desire to create systems that yield accurate data about teachers and that are perceived by teachers as fair and capable of providing meaningful information for professional growth. Emergent trends from the separate analyses illustrate the need to involve educators in evaluation system work and the necessity of reviewing teacher evaluation measures each year and across years.

The case studies in this project illustrate that reliable and valid measures can be identified and instruments can be created and vetted as components of teacher evaluation systems. In both DDOE and CCSD, hundreds of locally-generated student growth assessments have been designed and implemented, and have demonstrated acceptable reliability. Similarly, classroom observation measures have been constructed and deployed with increasing levels of rater consistency in FWCS and in CCSD. Central to the success of SLOs/SGOs and classroom observation rating improvements across all three agencies were two tenets. First, educator involvement in student growth measure design and in development of classroom observation instruments and protocols was essential for their acceptance by both teachers and administrators. And second, it was essential to routinely re-evaluate the status of measures.

The commitment of SDP Fellows and other staff members to monitor the ongoing progress of these measures ensured that data collected were accurate. In the initial stages, analyses resulted in rewriting assessment items and realizing the need for administrator training

in observation. As the new educator evaluation systems matured, more sophisticated analyses were undertaken to examine the relationships among measures and to consider whether factors beyond the measures were influencing scores. As these teacher evaluation systems continue to evolve, inclusion of educators in the process and targeted analytical projects will transform this work from a focus solely on the evaluation to a focus on robust and personalized professional learning for teachers.

In both FWCS and CCSD, teachers' overall effectiveness scores varied. The addition of a new measure and alterations to the student growth measure may have contributed to a downward trend and changes in overall evaluation scores in FWCS. The model in CCSD was stable across years and the general upward trend in overall scores may be attributable to teachers' familiarity with measures and their increasing reliability. It is advisable to examine teacher effectiveness scores across several years. Perhaps a longitudinal view would provide the best idea of teachers' impact compared to a look at only one year with one group of students.

A prominent takeaway from this project is the reinforcement that teacher evaluation must include multiple measures to create a comprehensive view of teacher effectiveness. For example, in the DDOE project, scores from Measure B assessments were used in conjunction with other data, not as a stand-alone measure. Likewise, given the disparity of correlational evidence between classroom observation scores and student growth measures in FWCS and CCSD, it is imperative that as many data points as possible be included to avoid one measure unjustifiably influencing teacher ratings, particularly when employment and compensation decisions rest on those data.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in Pittsburgh Public Schools*. Retrieved from ERIC database. (ED545232)

Danielson, C. (2011). Evaluations that help teachers learn. *The Effective Educator, 68*(4), 35–39. Retrieved from http://www.ascd.org/publications/educational-leadership/dec10/vol68/num04/Evaluations-That-Help-Teachers-Learn.aspx

Danielson, C. (2012). Observing classroom practice. *Educational Leadership, 70*(3). Retrieved from http://www.ascd.org/publications/educational-leadership/nov12/vol70/num03/Observing-Classroom-Practice.aspx

Donaldson, M. L. (2009). *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality*. Washington, DC: Center for American Progress. Retrieved from http://cdn.americanprogress.org/wp-content/uploads/issues/2009/06/pdf/teacher_evaluation.pdf

Dryfoos, J., Quinn, J., & Barkin, C. (Eds.). (2005) *Community schools in action: Lessons from a decade of practice.* Don Mills, Ontario, Canada: Oxford University Press.

Fullan, M., Hill, P., & Crevola, C. (2006) *Breakthrough.* Thousand Oaks, CA: Corwin.

Gullickson, A. R. (2009). *The personnel evaluation standards: How to assess systems for evaluating educators* (2nd ed.). Thousand Oaks, CA: Corwin.

Hanushek, E. A. (2007). The single salary schedule and other issues of teacher pay. *Peabody Journal of Education, 82,* 574–586. Retrieved from http://hanushek.stanford.edu/sites/default/files/publications/hanushek.2007%20PeabodyJEd%2082(4).pdf

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal, 6*(1), 73–112. doi:10.3102/0002831213517130

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Developing and selecting assessments of student growth for use in teacher evaluation systems (extended version)*. Los Angeles, CA: UCLA CRESST.

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 20*(1), 101–136.

Kane, M. (2013). Validating the interpretation and use of test scores, *Journal of Educational Measurement, 50*(1), 1–73.

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45*(1), 34–70. doi:10.1177/0013161X08327549

Koppich, J., & Rigby, J. (2009). Alternative teacher compensation: A primer. *Policy Analysis for California Education.* Stanford, CA: Policy Analysis for California Education. Retrieved from http://files.eric.ed.gov/fulltext/ED510159.pdf

Lemke, M., Thomsen, K., Wayne, A., & Birman, B. (2012). *Providing effective teachers for all students: Examples from five districts*. Washington, DC: U.S. Department of Education. Retrieved from https://www2.ed.gov/rschstat/eval/teaching/providing-effective-teachers/report.pdf

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Smith, M. S., & O'Day, J. (1990). Systemic school reform. [Special Issue]. *Journal of Education Policy, 5*(5), 233–267. doi:10.1080/02680939008549074

University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2015). Getting classroom observations right. *Education Next*, *15*(1), 63–68. Retrieved from http://educationnext.org/getting-classroom-observations-right/

Appendices

Delaware Department of Education

Ongoing Refinement Cycle

Subject	Refinement Year	Roll-Out Year
SS & Arts	14-15	15-16
CTE	15-16	16-17
Math, ELA, & Science	16-17	17-18
Languages & Other	17-18	18-19

Table X

Ongoing Assessment Refinement Cycle

Subject	Refinement Year	Roll-Out Year
Social Studies, Arts	2014–2015	2015–2016
Career Technical Education	2015–2016	2016–2017
Math, English Language Arts, Science	2016–2017	2017–2018
Languages, Other	2017–2018	2018–2019